# BAB II

# TINJAUAN PUSTAKA DAN DASAR TEORI

## 2.1 Tinjauan Pustaka

Untuk membedakan penelitian ini dengan penelitian sebelumnya maka penulis melakukan studi pustaka terhadap peneliti peneliti terdahulu.

Ardytha Luthfiarta, Junta Zeniarja & Abu Salam (2013). Peneliti melakukan peringkasan dokumen otomatis dengan menggunakan algoritma *Latent Semantic Analysis* (LSA). Berdasarkan percobaan-percobaan yang telah dilakukan dapat disimpulkan bahwa Peringkas Dokumen Otomatis dengan *Latent Semantic Analysis* (LSA) pada Proses Clustering Dokumen Teks Berbahasa Indonesia dapat meningkatkan kinerja clustering dokumen lebih baik dari pada Peringkas Dokumen Otomatis dengan Metode Fitur dan Proses Clustering Dokumen Standar, mengalami peningkatan dari tingkat akurasi 65,92 % untuk proses clustering standar menjadi 71,04% untuk proses clustering dokumen menggunakan peringkas dokumen otomatis dengan *Latent Semantic Analysis* (LSA).

Dimas Bagus C. W. (2017). Peneliti membuat aplikasi dengan metode metode text mining untuk melakukan analisis sentimen untuk menemukan topik topik yang terdapat pada data tweets dan melihat keterkaitan antar kata pada khasus pilkada DKI Putaran 2.

Keke Putri Utami (2017) melakukan pemodelan topik menggunakan *Latent Dirichlet Allocation* pada lima lokasi *tweet* di Kota Bogor dan rentang waktu tertentu, dan berhasil membentuk topik dengan informasi atau deskripsi topik

untuk setiap lokasi *tweet*. Deskripsi topik untuk setiap lokasi *tweet* menunjukkan topik-topik tersebut sedang banyak dibicarakan oleh pengguna Twitter pada rentang waktu yang ditentukan. Jumlah topik yang ditentukan untuk data csv setiap lokasi *tweet* telah dapat menghasilkan kumpulan kata yang membentuk topik dengan baik. Informasi yang mewakili isi topik dapat dimanfaatkan oleh pembaca atau *stakeholder* terkait dalam memahami setiap perkembangan isu terkini.

Nurina Savanti Widya Gotami, Indriati, Ratih Kartika Dewi (2018). Peneliti melakukan peringkasan teks otomatis secara ekstraktif pada artikel berita kesehatan berbahasa Indonesia dengan menggunakan LSA dapat diterapkan dengan cara LSA sebagai algoritme untuk mendapat kalimat kalimat yang memiliki keterkaitan kata dengan pendekatan secara semantik dengan menggunakan SVD sebagai fitur penghilang redudansi atau *noise* pada kata tertentu. Serta penggunaan *Cross method* LSA sebagai pengekstrasi ringkasan yang akan dipilih dari artikel berita kesehatan dalam data teks dokumen.

Tinjauan keenam ditulis oleh Septian Narsa Putra (2018). Peneliti membuat sistem yang dapat melakukan klasifikasi terhadap berita yang diupload pada akun twitter Divis Humas Polri ke dalam tiga katagori yaitu berita kegiatan polisi, komentar masyarakat dan layanan masyarakat selama empat tahun ke belakang. Setelah berita diklasifikasi kemudian dicari sentimen dari setiap topik. Metode yang digunakan adalah Naive Bayes Classifier.

Pada penelitian kali ini akan dibuat sistem yang dapat melakukan clustering berdasarkan kesamaan topik pada akun twitter pejabat publik, sehingga dapat

ditemukan topik yang tersembunyi dari tweets yang ada menggunakan *latent semantic analysis*.

Perbandingan dari penelitian-penelitian diatas dapat dilihat pada table 2.1.

**Tabel 2.1 Perbandingan Penelitian**

| Penulis | Objek | Metode | Hasil |
|---|---|---|---|
| Ardytha Luthfiarta, Junta Zeniarja & Abu Salam (2013) | Dokumen Teks Berbahasa Indonesia | *Latent Semantic Analysis (LSA)* | Sistem untuk proses c*lustering* dokumen teks berbahasa Indonesia. |
| Dimas Bagus C. W. (2017) | Data Twitter | Supervised Learning : Naïve Bayes dan Unsupervised Learning : Association Rule, K Means & Topic Modeling | Aplikasi dengan menggunakan metode metode text mining untuk melakukan analisis sentiment dan menemukan topik topik yang terdapat pada data tweets. |
| Keke Putri Utami (2017) | Data Twitter | *Latent Dirichlet Allocation* | Sistem pemodelan topik LDA pada 5 lokasi tweet di kota bogor membentuk informasi topik setiap lokasi tweet. |
| Nurina Savanti Widya Gotami 1, Indriati 2, Ratih Kartika Dewi 3 (2018) | Artikel berita kesehatan dalam data teks dokumen | *Latent Semantic Analysis (LSA)* | Sistem peringkasan teks otomatis secara ekstraktif pada artikel kesehatan berbahasa Indonesia. |
| Septian Narsa Putra (2018) | Data Twitter | Naive Bayes Classifier | Sistem klasifikasi berita pada akun twitter. |
| Diusulkan, Widya Sulistyani (2020) | Data Twitter | *Latent Semantic Analysis (LSA)* | Sistem yang dapat menemukan topik tersembunyi di dalam tweets menggunakan pemodelan topik Latent Semantic Analysis. |

## 2.2 Dasar Teori

### 2.2.1 Twitter

Twitter merupakan salah satu media sosial dengan layanan *microbloging* yang terkenal dan memungkinkan para penggunanya untuk menulis sesuatu atau yang biasa disebut *tweet*. Twitter digunakan untuk mengutarakan opini publik maupun berita resmi dari suatu instansi atau dari pejabat publik. Twitter dibangun oleh Jack Dorsey pada tahun 2006 dengan alamat http://www.twitter.com, jika seseorang ingin menggunakan twitter seseorang harus terlebih dahulu memiliki akun, untuk registrasinya dapat dilakukan pada alamat tersebut. Pengguna dapat menulis pesan berdasarkan topik dengan tanda #(tagar). Sedangkan untuk menyebut atau membalas pesan dari pengguna lain bisa menggunakan tanda @(diikuti nama akun yang akan dibalas).

### 2.2.2 *Text Mining*

*Text mining* memiliki definisi menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen, dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen.

### 2.2.3 Pemodelan Topik

*Topic Modeling* atau pemodelan topik merupakan metode *clustering* yang termasuk dalam *unsupervised learning*. Dalam *unsupervised learning* tidak ada label untuk suatu objek. Terdapat 3 tipe *clustering* yaitu *hard clustering*, *hierarchical clustering*, dan *soft/fuzzy clustering*. Pemodelan topik termasuk dalam *soft/fuzzy clustering* yang mana setiap objek dapat dimiliki lebih dari satu

*cluster* dengan tingkat tertentu. Sebagai contoh, perhitungan kemungkinan yang dimiliki objek untuk tergabung dalam suatu *cluster* adalah berbeda (Doig 2015).

Setiap harinya sejumlah besar data terkumpul dan menjadikan semakin banyaknya informasi yang tersedia. Pemodelan topik menyediakan suatu metode untuk mengatur, memahami, dan meringkas kumpulan dokumen. Pemodelan topik membantu dalam hal menemukan pola topik tersembunyi yang ada dalam kumpulan dokumen, memberikan keterangan dokumen sesuai dengan topik, memanfaatkan pemberian keterangan ini untuk mengatur, mencari, dan meringkas data teks. Pemodelan topik dapat digambarkan sebagai metode untuk menemukan kelompok kata (topik) dari kumpulan dokumen yang dapat merepresentasikan dengan baik informasi yang ada dalam kumpulan dokumen tersebut (Nair 2016). Terdapat beberapa teknik yang dapat digunakan untuk pemodelan topik, salah satunya adalah pemodelan topik *Latent Semantic Analysis*.

### 2.2.4 Latent Semantic Analysis

LSA *(Latent Semantic Analysis)* adalah metode statistik aljabar yang mengekstrak struktur semantik yang tersembunyi dari kata dan kalimat, untuk mencari interelasi diantara kalimat dan kata, digunakan metode aljabar *Singular Value Decomposition* (SVD). SVD adalah cara dekomposisi matriks yang digunakan untuk mencari kesamaan antar segmen kata**.** SVD adalah komponen pemrosesan yang mengkompresi informasi yang berkaitan dalam jumlah besar ke dalam ruang yang lebih kecil. Proses awal dari LSA adalah merepresentasikan isi kata dalam matriks dua dimensi yang besar yang berisi *bag-of-words* dari tweets dimana kolom merepresentasikan tweet, dan baris mewakili kata/istilah.

Disamping mempunyai kapasitas relasi model diantara kata dan kalimat, SVD ini mempunyai kapasitas reduksi noise yang membantu untuk meningkatkan akurasi. Formula dari SVD adalah sebagai berikut :

$$A = USV^T$$

Keterangan :

A : matriks asal

U : *orthonormal eigenvector* dari $AA^T$

S : matriks diagonal

$V^T$ : *transpose* dari *orthogonal* matriks V

Dekomposisi ini memungkinkan dimensi matriks asal untuk dilakukan reduksi dimensi. Dengan proses reduksi dimensi terhadap perkalian matriks SVD, maka akan diperoleh penyederhanaan dan pembobotan dari matriks asal dengan mengambil sebagian besar dari struktur penting antara kata kunci dengan kalimatnya.